A Time-to-Event Analysis of Heart Failure via Electronic Health Records

Austin Falco, Aidan Kearns, Miranda Riggs, Tahmeed Tureen

Statistics 404, Fall 2017

December 12, 2017

EXECUTIVE SUMMARY

This research project seeks to learn more about the risk factors that increase the likelihood of developing heart failure and how much of an impact these risk factors have. It also looks to assess the time until a heart failure diagnosis occurs given a certain set of characteristics for a person. The data for this project is provided by the Michigan Genomics Initiative (MGI) from the University of Michigan. The data is electronically stored medical data called Electronic Health Records (EHR).

The results of the study conclude that males are affected by cardiomyopathy, severe chronic kidney disease, hypertension, and age, while females are affected by Type II diabetes, sleep disorder, hypertension, and age. These findings are supported by similar conclusions drawn in readings from the Centers for Disease Control and Prevention (CDC) and the National Institute of Health (NIH).

INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC), heart failure occurs when the heart cannot pump enough blood to support the body's organs. About 50% of the people who are diagnosed with heart failure die within the first five years of their diagnosis. This makes any study on heart failure and its risk factors immensely valuable. In our study, we attempted to research this disease and its risk factors by using Electronic Health Records (EHR) from the University of Michigan Health System. We defined two research goals:

1. Estimate and compare the time until a heart failure diagnosis occurs given a certain set of characteristics for a person.
2. Evaluate how certain risk factors affect the time until a heart failure diagnosis occurs.

Essentially, we want to understand which risk factors have the potential of causing heart failure and how much impact they may have on the risk of heart failure.

CONCLUSIONS

Based on the findings drawn from the statistical analysis in this project, there exists four significant risk factors that affect males and four significant risk factors that affect females. The males in our study were affected by cardiomyopathy, severe chronic kidney disease, hypertension, and age. The females in our study were affected by Type II diabetes, sleep disorder, hypertension, and age. Upon further review, we note that the results of our research are consistent with readings from the CDC and the National Institute of Health (NIH).

METHODS

Survival Analysis was used was used to estimate and compare heart failure over time for the groups of interest. Survival analysis is a set of methods for analyzing data where the outcome variable is the time before an event of interest occurs. This event could be something like death, occurrence of a disease, or a mechanical failure. In this study, the event of interest is the amount of time until a person is diagnosed with heart failure.

Survival analysis aims to estimate the proportion of a certain group that will survive past a given time. In our case, the number of interest is the proportion of patients who will go without a heart failure diagnosis after 5, 10, 15 years, etc. In addition to estimating survival rate, survival analysis can also be used to assess how much risk factors influence the potential of receiving a heart failure diagnosis.

The specific survival analysis method used in our study is called a Cox Proportional Hazard (CPH) model. This method works well with quantitative variables like age and weight. It can also consider more than one covariate in the analysis. In our situation, this means instead of just observing survival for similar 50-year-old men, CPH takes into account covariates like arrhythmia, cardiomyopathy, and Type II diabetes. It estimates differences in survival between these covariates. These estimates come in the form of a hazard ratio. This ratio is the risk of failure, heart failure in our case, for one group compared to another. For example, if the hazard ratio for sex is 1.6 for males versus females, then males are 1.6 more likely to experience heart failure at any given time.

ANALYSIS AND RESULTS

This study finds the following hazard ratios for the risk factors included in **Table 1** and **Table 2** by using the Cox Proportional Hazards model.

In the context of this research question, the interpretation of these hazard ratios can be best explained through an example. Consider two individuals with the same characteristics, the same diagnoses, and all else the same except individual A has cardiomyopathy, and individual does not. If the hazard ratio for cardiomyopathy is 10.203, this means that individual A (the one with cardiomyopathy) is 10.203 times more likely to experience heart failure than individual B(the one without cardiomyopathy).

Survival analysis was performed on both the male and female groups to find the hazard ratios for each of the risk factors.. The hazard ratios for each risk factor are as follows:

TABLE 1. Hazard ratios for male cohort

Risk Factor	Hazard Ratio
Cardiomyopathy	10.203*
Severe Chronic Kidney Disease	3.08*
Hypertension	1.337*
Age	1.052*
Type 2 Diabetes Mellitus	3.526
Hepatitis C	1.945
Moderate Chronic Kidney Disease	1.725
Emphysema	1.684
Arrhythmia	1.259
Failure Chronic Kidney Disease	1.191
Morbid Obesity	1.064
Sleep Disorder	1.036
Type 2 Diabetes Mellitus*Age	0.987
Hypercholesterol	0.911
Mild Chronic Kidney Disease	0.569

Source: Michigan Genomics Initiative, University of Michigan
Note: n = 4204
*p < .05

TABLE 2. Hazard ratios for female cohort

Risk Factor	Hazard Ratio
Type 2 Diabetes Mellitus	9.350*
Hypertension	1.442*
Sleep Disorder	1.764*
Age	1.052*
Failure Chronic Kidney Disease	4.179
Hepatitis C	3.474
Moderate Chronic Kidney Disease	2.169
Severe Chronic Kidney Disease	2.070
Mild Chronic Kidney Disease	1.675
Morbid Obesity	1.635
Arrhythmia	1.563
Emphysema	1.333
Type 2 Diabetes Mellitus*Age	0.973
Cardiomyopathy	0.967
Hypercholesterol	0.776

Source: Michigan Genomics Initiative, University of Michigan
Note: n = 4929
*p < .05

For the male cohort, this study shows that cardiomyopathy, severe chronic kidney disease, hypertension, and age are the the most significant risk factors for heart failure. This can be seen in **Table 3** below.

TABLE 3. Significant hazard ratios for male cohort

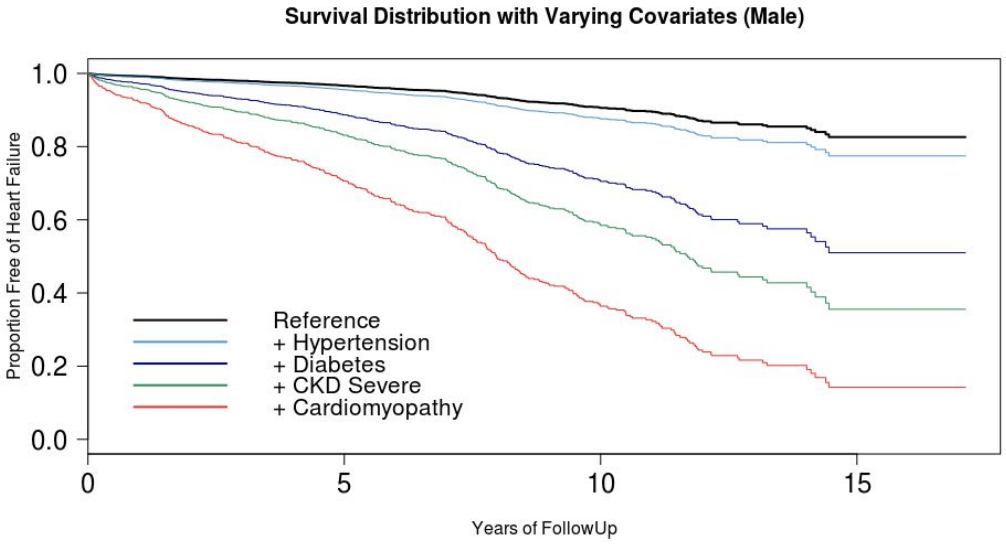
Risk Factor	Hazard Ratio
Cardiomyopathy	10.203
Severe Chronic Kidney Disease	3.080
Hypertension	1.337
Age	1.052

Source: Michigan Genomics Initiative, University of Michigan

Note: n = 4204

Note: Significant at 5% significance level

Cardiomyopathy has the largest hazard ratio out of all the significant risk factors with a hazard ratio of 10.203. This suggests that cardiomyopathy has the largest effect on the risk of heart failure for the males. The effects of each individual risk factor can also be seen in **Figure 1** below.



Source: Michigan Genomics Initiative, University of Michigan
Note: n = 4204
Note: Baseline patient is 60 years old and has been diagnosed with Arrhythmia, Mild CKD, and Hypercholesterol

Figure 1. Survival Distribution for Males

For the female cohort, this study shows that Type II diabetes mellitus, sleep disorder, hypertension, and age are the most significant risk factors for heart failure. This can be seen in **Table 4** below.

TABLE 4. Significant hazard ratios for female cohort

Risk Factor	Hazard Ratio
Type 2 Diabetes Mellitus	9.350
Sleep Disorder	1.764
Hypertension	1.442
Age	1.052

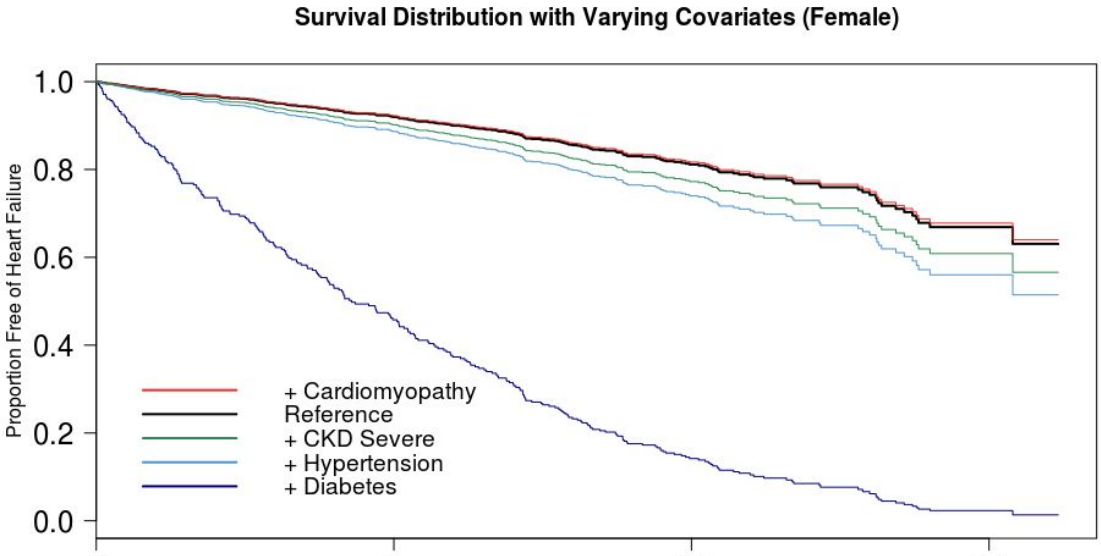
Source: Michigan Genomics Initiative, University of Michigan

Note: n = 4929

Note: Significant at 5% significance level

Note: Type 2 Diabetes Mellitus had a 95% confidence interval of (1.433, 61.024)

The risk factor of special interest is Type II diabetes. This risk factor has a 95% confidence interval of (1.433, 61.024) for the hazard ratio. The confidence interval can be interpreted as a range of values that contains the "true" hazard ratio for Type II diabetes. This range of values is of interest due it spanning a large range of numbers. Additionally, the values contained in this range suggest that Type II diabetes has a very large effect on the risk of heart failure for females. Overall, this range of values suggests that although Type II diabetes has a large effect on females, the effect itself can be as small as 1.433 or as large as 61.024. This irregularity may be a cause of the EHR data itself, which is explained more in depth in the **limitations** section.



Source: Michigan Genomics Initiative, University of Michigan
Note: n = 4929
Note: Baseline patient is 60 years old and has been diagnosed with Arrhythmia, Mild CKD, and Hypercholesterol

Figure 2. Survival Distribution for Females

This exaggerated effect on the risk of heart failure for females can be seen in **Figure 2** above, along with the other significant risk factors. An interesting trend to note is that cardiomyopathy for females seems to have a positive effect. This surprising result may be due to study limitations (such as the large amount of missing data or sample size limitations), or this result could be true. Future research should be conducted to further analyze these risk factors.

ASSUMPTIONS

Several assumptions were made before the analysis was conducted. First, the risk factors (covariates) for heart failure used in our CPH model were chosen based on readings from the

CDC and American Heart Association (AHA). According to the readings, the variables used in our study appear to be the leading causes of heart disease. We made the assumption that these conditions would be the best predictors for heart failure, also.

A goal of the study was to make conclusions on conditions that impact the likelihood of heart failure for the general population of the United States. However, the study only uses data from the University of Michigan Health System. So, we make the assumption that this data is representative of the general United States population. In addition, we are assuming that the diagnosis made in the University of Michigan EHR are all accurate. When a doctor diagnoses a patient with a disease, it is recorded into the EHR. If someone in the data was diagnosed with hypertension it is assumed that they have hypertension and the diagnosis is not false.

Finally, the CPH model used in the analysis requires the assumption that the effects of predictor variables on survival are constant over time. In other words, all of the covariates for heart failure have the same effect on likelihood of survival at any given time. For example, Type II diabetes has the same effect on the chance of heart failure 5 years after diagnosis that it does 10 years after diagnosis.

DATA

The data set is a collection of electronically stored information regarding patients from the University of Michigan Health System. These data entries are called Electronic Health Records (EHR). For this study, the EHR were provided by the Michigan Genomics Initiative (MGI) at the

University of Michigan, Ann Arbor. The data represents approximately 1800 patients and consisted of approximately three billion data entries. Each entry is a diagnosis for *one* disease that was recorded and stored electronically into a database by the health system. The data contains personal information about each patient such as their age, weight, sex, diseases, and illnesses contracted. Therefore, the data is confidential and non-accessible to the general public.

It is important to note that EHR are not intended for research purposes, so doctors or nurses are not required to fill in the EHR entirely. This means that EHR are prone to large amounts of missing data. In this study, there were multiple counts of missing data (note that the exact amount of missing data was not originally recorded.) The initial EHR was provided in three separate datasets which were matched using patient ID's into a single dataset for further analysis.

LIMITATIONS

The limitations of this research project were due to the behavior and format of the EHR itself. These limitations influenced how the research project was conducted and has the potential of affecting the results drawn from the research.

As noted in the **data** section, the EHR are not collected for research purposes. Therefore, there is no standard protocol or format for the way the data is collected. This leads to textual differences in diagnoses. For example, a nurse may diagnose a patient with malignant hypertension and record this as "Hypertension, Malignant" while a different nurse could enter "M. Hypertension"

into the database. Therefore, the hospital diagnoses could not be used to match the data when creating an analysis dataset (the study cohort, or the group of patients chosen for this study).

To tackle to this limitation, the UM Health System's diagnosis codes were used. However, diagnoses codes changed after 2014 which also presented some difficulty in terms of defining a cohort. For example, before 2014, hypertension had two separate codes for malignant hypertension and benign hypertension. After 2014, there was only one. This sounds relatively easy to work around, however, some of the other covariates had multiples codes before 2014 which were collapsed to multiple codes as well. This made the data manipulation process difficult.

Finally, the EHR is a non-continuous stream of data with very little contextual information about the patients. There are three different scenarios of this, which are explained below:

(1) A patient has all of their diagnoses stored at the same health system. All of their diagnoses are stored in a single database.
(2) A patient received one or two diagnoses at a different health system. Not all of their diagnoses are stored in a single database.
(3) A patient does not necessarily have a "final" diagnosis. This is because the final diagnosis in the health system does not indicate whether or not the patient will back for another diagnosis in the future.

These three behaviors have the potential of influencing the results in our study because they do not provide all of the necessary contextual background for the researcher to make concrete conclusions about a specific patient.

COHORT DEFINITION

We define the sample of people who were analyzed in this research as the cohort. The cohort was decided based on several filtration guidelines which we defined after considering all of the potential limitations in the project. First and foremost, all patients who were under the age of 39 were dropped from the the study. This was done because having heart failure before the age of approximately 40 years is considered an abnormality. Secondly, patients with less than two years of EHR data were dropped. We chose to do this because we believe that patients with less than two years of data are not helpful for the survival analysis method used in this project. Finally, any patient (age > 39) who was diagnosed with heart failure in their second diagnosis was dropped. This was done because a total of two diagnosis is not sufficient for survival analysis to draw any meaningful result. Following these guidelines, a final cohort of 9133 patients was defined. There are 4929 females and 4204 males in the cohort.

NEXT STEPS

In the future, shrinkage methods could be used to learn more about the risk factors in this study. In general, shrinkage methods are a way to use new analyses to go more in depth and learn more about the effects of certain variables – in this case, risk factors for heart failure. In particular, this study could benefit from performing shrinkage methods using the significant factors for males

and females. Using shrinkage methods could help researchers learn more about the impact these risk factors have on heart failure.

Additionally, further research could be performed on the significant factors found in this study (which were cardiomyopathy, severe chronic kidney disease, hypertension, and age for males, and Type II diabetes mellitus, sleep disorder, hypertension, and age for females). This study would benefit from further research on the behaviors of these risk factors and how they interact with heart failure. Researchers could also look into possible ways of preventing the development of these risk factors, which would possibly reduce the risk of heart failure.

Another opportunity for future research is looking to the research potential of EHR. As EHR data has not previously been used for research purposes, further research into understanding how EHR works and how it can be used for research purposes could result in a wealth of knowledge. As there are large amounts of data contained in EHR, many new insights could be obtained by understanding and taking advantage of EHR in the future.

ACKNOWLEDGEMENTS

The authors of this report would like to acknowledge the following people for their contribution to the overall success of the research project:

Researchers:

(1) Deesha Bhaumik, Undergraduate Student, Johns Hopkins University

(2) Christian Erickson, Undergraduate Student, Hope College, Michigan

(3) Tahmeed Tureen, Undergraduate Student, University of Michigan

Supervisors:

(4) Dr. Philip Boonstra, Assistant Professor, Department of Biostatistics, University of Michigan

(5) Dr. Zhenke Wu, Assistant Professor, Department of Biostatistics, University of Michigan

(6) Dr. Matthew Zawistowski, Research Specialist, Department of Biostatistics, University of Michigan

The authors of this report would also like to acknowledge Professor Mary Ann Ritter of the Department of Statistics, University of Michigan for her guidance in the writing and presentation of this research project.